Title	強化学習と脳における報酬系の情報処理(脳化学2,数学者のための分子生物学入門-新しい数学を造ろう-)
Author(s)	石井, 信; 柴田, 和久
Citation	物性研究 (2006), 87(3): 467-472
Issue Date	2006-12-20
URL	http://hdl.handle.net/2433/110688
Right	
Type	Departmental Bulletin Paper
Textversion	publisher

強化学習と脳における報酬系の情報処理

奈良先端科学技術大学院大学情報科学研究科　　石井　信

レクチャーノート作成：柴田　和久（奈良先端科学技術大学院大学）

1．はじめに

行動と結果の連合（ある行動に対する結果が満足のいくものならば、行動と結果の結びつきが強まる）は、動物の行動学習の基本だと考えられてきた。例えばラットが壁にあるボタンを押した（行動）ときにチーズ（報酬）を与えたり電気刺激（罰）を与えたりすることで、ラットはボタン押しと結果の結びつきを学習する。これは行動心理学では「効果の法則」と呼ばれるもので、自律的な学習であり、明示的な教師の必要はない。

近年活発に研究されている強化学習は、このような学習の枠組みに行動系列の考え方を取り入れ、将来にわたる報酬の和を最大にするような行動系列を獲得するための学習法である。本稿では、まず強化学習の考え方を説明し、次に脳における報酬系の情報処理と強化学習の関係について紹介する。

2．強化学習の考え方

強化学習の要素は、状態の価値、それに基づく行動決定法、行動に対する報酬に分けることができる。学習者の課題は、ある行動に対して得られる報酬を手がかりにして、状態の価値を学習する（状態価値関数の学習）、そして現在の状態価値から最適な行動決定法を決める（方策の学習）という２つのステップを繰り返すことである。学習過程で正しい行動を明示的に指示されることはなく、行動に対する報酬のみに基づいて学習が進む。これは下図のように描ける。ここで s_t は現在の環境の状態で学習システムに対する入力である。一方 u_t は学習システムが環境の状態 s_t に対して取った行動であり、学習システムの出力である。また r_{t+1} は行動 u_t に対する報酬を意味する。環境は u_t に応じて新たな状態 s_{t+1} に遷移する。方策は s_t から u_t への写像になる。

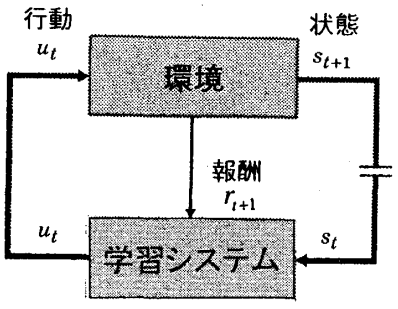

図１.強化学習の枠組み

ここでの目的は、将来にわたる累積報酬を最大にすることである。累積報酬 R(t)は以下の式で与えられる：

$$R(t)=r_{t+1}+\gamma r_{t+2}+\cdots\gamma^k r_{t+k+1}=\sum_{k=0}^{T}\gamma^k r_{t+k+1} \qquad (1)$$

r_{t+1}は時刻 t の行動に対する即時報酬を表し、T は行動の終了時刻を表す。γは遠い将来の即時報酬ほど割り引いて考えるための割引率（discount factor）で、$0\leqq\gamma\leqq1$ である。

2．1　TD 学習

状態遷移にマルコフ性を仮定すると、現在の状態と行動から次の時刻の状態と報酬を予測することができる。さらに繰り返し計算により、すべての将来の状態と報酬を予測することができる。ここで、現在の状態がどれくらい良いのかを計るものとして、状態価値関数 V(s)を考える。良さは将来にわたって得られる報酬の期待値として定義する。また π(s,u)を状態 s のもとで行動 u を取る方策とする。方策 π のもとで状態 s の価値 V(s)は以下のように書ける：

$$V^{\pi}(s)=E_{\pi}\{R_t\mid s_t=s\}=E_{\pi}\left\{\sum_{k=0}^{T}\gamma^k r_{t+k+1}\mid s_t=s\right\} \qquad (2)$$

この式から V(s)は再帰的な構造を持つことがわかる。すなわち、

$$V^{\pi}(s_t)=r_{t+1}(s,\pi(s))+\gamma V^{\pi}(s_{s+1}) \qquad (3)$$

である。もっとも簡単な場合での V(s)の学習は、s_tで行った行動に対する報酬 r_{t+1}を観測し，（3）式の左辺と右辺の差を最小にするよう V(s)を更新するやり方である。このときの差 δ_tを TD(Temporal Difference)誤差と呼ぶ：

$$\delta_t=r_{t+1}+\gamma V^{\pi}(t+1)-V^{\pi}(t) \qquad (4)$$

この TD 誤差を使った V(s)の更新アルゴリズム（TD 学習）の例を以下に示す：

①V(s)を適当に初期化

②各試行（エピソード）の度に以下を繰り返す

i.状態を初期化し s_0 とする

ii.s_t が終了状態になるまで以下を繰り返す

a.価値の高い状態へ導く行動 u_t を選択

b.行動 u_t を実行、次の状態 s_{t+1} と即時報酬 r_{t+1} を観測

c.TD 誤差 δ_t を計算する

$$\delta_t=r_{t+1}+\gamma V(s_{t+1})-V(s_t)$$

d.状態価値関数を以下で更新

$$V(s_t)\leftarrow V(s_t)+\alpha\delta_t$$　（α は小さい正の値）

2．2　Q学習

各状態において、最適な行動を取った際の行動価値関数の学習（同時に最適な方策の学習を行っていることになる）をQ学習と呼ぶ。ここで行動価値というのは、ある方策のもと状態sに対して行動uを取ることの価値で、以下のように定義できる：

$$Q^{\pi}(s,a)=E_{\pi}\{R_t \mid s_t=s,u_t=u\}=E_{\pi}\left\{\sum_{k=0}^{T}\gamma^k r_{t+k+1} \mid s_t=s,u_t=u\right\} \qquad (5)$$

Q学習の更新アルゴリズムの例を以下に示す：

①Q(s,u)を適当に初期化

②各試行（エピソード）の度に以下を繰り返す

i.状態を初期化しs_0とする

ii.s_tが終了状態になるまで以下を繰り返す

a. 以下の式で行動u_tを選択する。時々ランダムな行動も入れる

$$u_t=\arg\max_u\left(Q(s_t,u)\right)$$

b.行動u_tを実行、次の状態s_{t+1}と即時報酬r_{t+1}を観測

c.TD誤差δ_tを計算する

$$\delta_t=r_{t+1}+\gamma\arg\max_u(Q(s_{t+1},u))-Q(s_t,u_t)$$

d.行動価値関数を以下で更新

$$Q(s_t,u_t)\leftarrow Q(s_t,u_t)+\alpha\delta_t \qquad (\alpha\text{は小さい正の値})$$

2．3　アクター・クリティック

アクター・クリティック法は、価値関数とは独立に、方策を表現するモジュールを別に持つ学習法である。方策を管理するモジュールをアクター（actor）、そして価値関数を学習するモジュールは、アクターの行動に批評を行うためにクリティック（critic）と呼ばれる。状態価値関数、方策の両方ともに、同じ状態価値関数のTD誤差を用いて更新される。

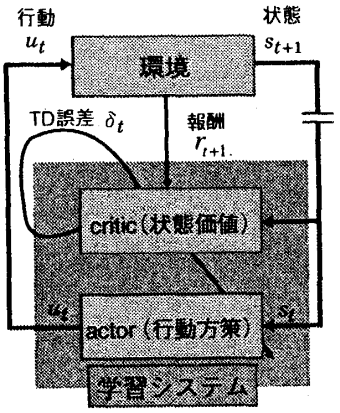

図2．アクター・クリティック法

アクター・クリティック法の更新アルゴリズムの例を以下に示す：

①V(s)を適当に初期化
②各試行（エピソード）の度に以下を繰り返す
　i.状態を初期化し s_0 とする
　ii.s_t が終了状態になるまで以下を繰り返す
　　a.アクターにおいて以下の確率で行動 u_t を選択

$$P^{\pi}(u \mid s_t) \propto \exp[\beta m(s_t, u)] \qquad (\beta \text{はある正の値})$$

　　b.行動 u_t を実行、次の状態 s_{t+1} と即時報酬 r_{t+1} を観測
　　c.TD 誤差 δ_t を計算する

$$\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t)$$

　　d.クリティックにおいて状態価値関数を以下で更新

$$V(s_t) \leftarrow V(s_t) + \alpha\delta_t \qquad (\alpha \text{は小さい正の値})$$

　　e.アクターにおいて効用関数を以下で更新

$$m(s_t, u_t) \leftarrow m(s_t, u_t) + k\delta_t \qquad (\text{k は小さい正の値})$$

3．脳における報酬系の情報処理と強化学習

3．1　ドーパミン経路と強化学習

Olds と Milner は 1954 年に、ラットの辺縁系に電極を指しておいてラットがレバーを押すことで電流が流れるようにしてやると、ラットはレバーを次々に押して自分自身を刺激するようになることを発見した。Olds と Milner は、これは辺縁系を刺激するとラットが快感を覚えるためと解釈し、辺縁系が脳における報酬処理に関わることを見出した。辺縁系の中でもドーパミン経路は報酬系と特に関わりが深いとされている。ドーパミンは腹側被蓋や黒質緻密部から、扁桃体、側坐核、前頭前野、線条体に投射されており、腹側被蓋から側坐核へのドーパミン投射をブロックすることで、上記の自己刺激が起こらなくなることが確認されている。また Schultz らは腹側被蓋のドーパミンニューロンが、条件付けにおいて TD 誤差を表現するような振る舞いを見せることを発見した。この実験では、サルに光刺激のあとにレバーを押すとジュースによる報酬がもらえることを学習させる。そのときのドーパミンニューロンの発火系列を眺めると、学習のはじめと終わりでは、その特徴に変化が現れるのがわかる（図３）。学習初期のドーパミンニューロンはジュースの報酬そのものに反応するが、後期は光刺激そのものに大きく反応するようになる。さらに面白いことに、このとき報酬を与えないと、報酬が出る時間にドーパミンニューロンの発火率が減少する。この活動変化は TD 誤差でよく説明でき、Fiorillo らの実験では、報酬確率に対して発火確率がシステマチックに増減することなども確認されている。

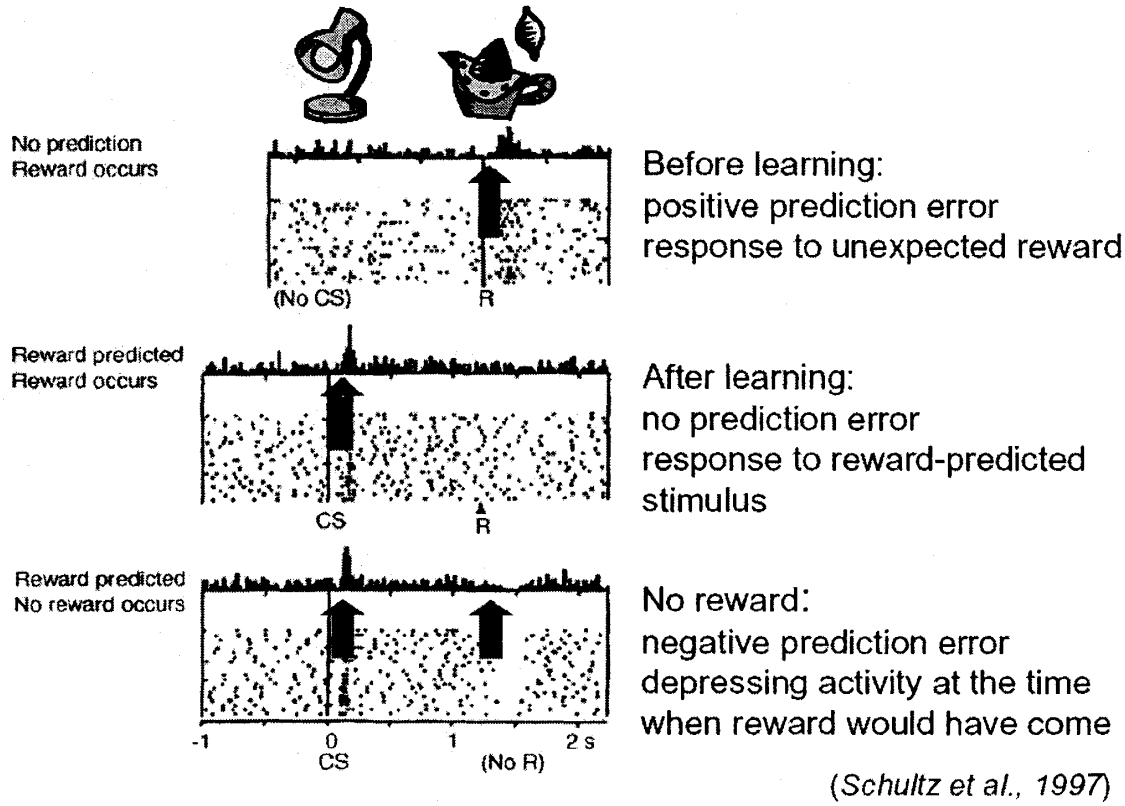

図３．Schultz の実験

３．２　線条体ニューロン活動の報酬依存性

Schultz らの実験を契機に、大脳基底核のニューロンは報酬の情報処理に関係するという観点で研究が行われた。設楽らは、側坐核のニューロンが報酬をもらえるまでの時間に応じて応答を変化させることを見出した。また川越らは、上下左右どれか 1 つのダーゲットが点灯し、その 1 秒後にターゲットに正しく眼を向けるとジュースの報酬がもらえるという課題を行っているサルの、線条体尾状核のニューロンの活動を計測した。尾状核のニューロンの多くは、特定の方向への眼球運動に先立って活動したが、その強さは、成功した場合にもらえるジュースの有無によって大きく変化した。この結果は、線条体のニューロンは単に運動そのものを表しているのではなく、運動の結果得られる報酬を予測しているのではないかということを示唆している。

３．３　大脳基底核における強化学習モデル

これまで紹介してきた知見から、大脳基底核の回路は TD 誤差を使った強化学習を行っているのではないかという仮説が提案されている。大脳基底核は線条体（Striatum）や黒質、淡蒼球、視床、視床下核で構成されており、興奮性/抑制性が入り混じる複雑な投射関係を持っている（図５）。

Barto らは、基底核の回路がアクター・クリティックの強化学習を実現しているという仮説を打ち出した。線条体のマトリックスはアクターとして、GPi/SNr を通じて脳幹や

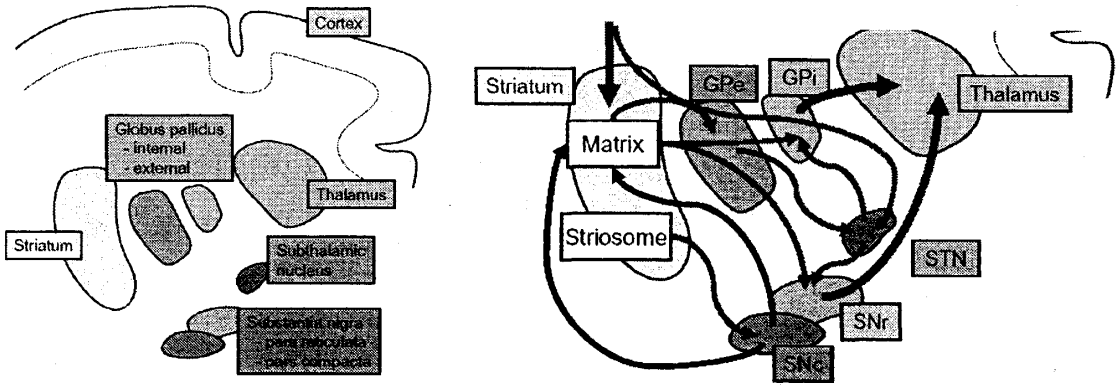

図４．大脳基底核の構成（左）と投射関係（右）

大脳皮質の行動出力を選択する。一方線条体のストリオゾームはクリティックとして報酬予測を行い、その投射先のドーパミンニューロンで TD 誤差が計算され、その線条体への投射によりアクター・クリティックの学習が行われる。O'Doherty らの行った fMRI 実験からは、線条体の腹側部がアクターとして、背側部がクリティックとして働くという仮説を提案しているが、線条体の腹側部ほどストリオゾームの占める割合が高く、逆に背側部ほどマトリックスの占める割合が高いことを考えると、この結果と Barto らのモデルはつじつまが合っている。

一方、銅谷らは、線条体で報酬予測と行動選択のすべて行われるのではなく、ストリオゾームでは状態価値関数が、マトリックスでは行動価値関数が学習されるという仮説を提案している。実際の行動選択はマトリックスの下流の GPi/SNr で、あるいはそれを含む大脳皮質－基底核ループのダイナミクスの結果として起こる、という可能性がある。

３．４　前頭皮質と環境モデル

大脳基底核で表現される強化学習は、観測だけから報酬予測を行う、モデルフリーの強化学習とされている。一方、部分観測などで環境の隠れ状態を推定するようなモデルベースの強化学習は、前頭前野で行われていると考えられている。近年環境のモデルの同定が必要な強化学習課題を用いた fMRI 実験なども行われている。

参考文献

大脳基底核と報酬予測, 銅谷賢治, 数理科学, No.512, February 2006, pp. 69-76

付録

本稿で紹介している TD 学習を、迷路課題を例に取り matlab によって実装しました。説明書（instruction.pdf）および matlab プログラム（demo_maze.m）は下記 url から入手できます。

http://www.cns.atr.jp/~kazuhi-s/works/RL/demo_maze.html